Workshop Discussion Report

NCMC-12: Data Acquisition, Handling, and Visualization

Goals of the Discussion

- Identify key informatics and data infrastructure needs of the Practitioners of high-throughput materials research.
- Summarize, prioritize, and communicate these needs to Suppliers of instruments and software.

Discussion Format

- Breakout Discussion Sessions, with 3 groups rotating through 3 topic areas.
- NIST moderators and discussion leaders.

Session #1: Automation, Integration and Central Database Tools
Moderator: Michael Fasolka, Director, NIST Combinatorial Methods Center
Notes: Matt Becker, Polymers Division, NIST
Carol Laumeier, NCMC, Polymers Division, NIST

Guiding Questions
***Practitioners**:*

- Instrument Automation and Integration
  a) What are the barriers you face in *automating* custom built instruments for high-throughput operation?
  b) What are the barriers you face in *automating* purchased instruments for high-throughput operation?
  c) What barriers do you face in integrating custom built instruments into a high-throughput workflow?
  d) How can the instruments you own now be improved in terms of automation and workflow integration?
  e) What instruments would you like to automate/integrate today?
- Databases
  a) Is your central database custom built or purchased? What platform?
  b) What kinds of data does your database accommodate already? What does it need to accommodate that it does not now?
  c) What data format do you use for interoperability? Custom? Other, e.g. XML?

- ***Suppliers**:*
  In addition to Practitioner Questions, consider:
  a) What data formats and database platforms are your instrument/software compatible with?
  b) Can one customize your software/instrumentation for automation/interoperability?
  c) Would you consider providing open source interoperability for your instruments/software?
  d) What is the main barrier to automating your instrumentation?
  e) What is the main barrier to providing instrument/software interoperability?

Discussion Summary

Barriers to instrument automation and integration

- Several participants named this as the single biggest problem in the development of high-throughput workflows. All participants named this to be a severe barrier to accomplishing high-throughput research.
- Practitioner participant companies report severe barriers to both automation and integration of instruments and software into high-thoughput workflow systems. The cost of these data barriers is high. One participant noted that the cost of integrating an instrument into an informatics system can be more than the original cost of the equipment, and could represent person-years of labor. Another participant noted that integration and automation issues formed a barrier to combi and high-throughput innovation; since instrument additions a workflow are expensive to accomplish, there is a "large barrier to starting something new."
- Software and instrument supplier participants note a fractured marketplace for informatics products, and state difficulty in determining common needs from the large variety of customers they serve. In addition, these parties state that it is difficult to strike a balance of flexibility in their products. For example, open source automation software can leave an instrument open to damage by inexperienced users. The suppliers state a willingness to provide data tools that will ease automation and integration, but note a lack of solid common targets to address. Some suppliers already provide output in open source formats, and recognize the need for flexibility in device and software design for combi and high-throughput practitioners.

Challenges and Opportunities

- Several participants reported that they would currently choose one instrument over another if it included flexible, open source software that eased integration and automation. Indeed, many participants would rank instrument interoperability as the most important factor in instrument performance. In addition, easy integration was rated by most participants to be more important than price. One participant stated that his company would be willing to pay an additional 100% of an instrument's cost if it included adequate informatics integration features.
- Practitioner participants identify the lack of common or standard interface data formats as the biggest challenge to integration. They note that some suppliers are now providing data output into XML format, and that based on widespread use and its flexibility XML seems to be a leading contender for an interface data format standard. However, they note a lack of XML schema appropriate for their work, and where schemas exist, they inadequately describe the data sets they most commonly use. Indeed, one participant noted that the data problem faced by combi practitioners is closer to the manufacturing sector (which has strong interoperability standards in place) than traditional R&D, and that the problem should be approached with a manufacturing mindset. Overall, there is an opportunity for suppliers to cooperate on data output formats and help identify the kinds and quality of schema they can commonly provide customers with. Practitioners state that a standard data format should be a) ascii based and b) well documented with metadata tags.
- The biotechnology sector is well ahead of materials research sectors in solving these issues. This includes practitioners who are generally focused on exactly what they need

from informatics interoperability and a supplier community that is better geared to providing users with instruments and software that help ease interoperability and system integration.

- Practitioner participants identify open source software as another high priority that would lower the barriers to workflow integration and automation. Many companies state that they would pay a premium for open source software if it was provided with an instrument. Others stated that they would assume risk of instrument damage if the instruments were provided with suitably flexible software packages that eased integration and automation. Some companies expressed a willingness to share software improvements they accomplished on open source code. In every case, it was stated that provided software must be very well documented if it is to be useful.

Session #2: Data Analysis, Mining, and Visualization
Moderator: Kirsten Genson, NCMC, Polymers Division, NIST
Notes: Leah Lucas, NCMC, Polymers Division, NIST

Guiding Questions
Practitioners:

- Automated Data Analysis
  a) What kinds of data do you perform routine automated analysis on now? What platforms do you use to do this?
  b) What is your biggest need for automated data analysis now? How would this improve your work?
  c) What might be your biggest automated data analysis need in 5 years?
  d) How important is image analysis to your work? Automated image analysis?
  e) How important is spectral analysis to your work? Automated spectral analysis?
- Data Mining
  a) Where do you employ datamining techniques today? For analyzing combinatorial library data? For analyzing databases of literature data? What kinds of data?
  b) What kind of commercial datamining tools, if developed, would you use? How would these tools improve your work?
- Data Visualization
  a) Do you use software tools to visualize large, multivariate data sets? How do you use these tools? What platforms do you use?
  b) What kinds of data would you like to visualize?
  c) What is your biggest need in terms of data visualization? If you had this tool, how would it improve your work?

Suppliers:
In addition to Practitioner Questions, consider:

a) What kinds of automated data analysis capabilities (including image data) are built into your instrument's software package?
b) If your instrument and/or software does not enable automated data analysis, can it be modified to do so?
c) What is the main barrier to providing automated data analysis routines? Lack of good algorithms? Lack of a suitable market for these tools?

Discussion Summary

Automated Data Collection and Analysis and Platforms Used:

- **Scope of needs:** Participant companies report the need to perform automated data analysis in nearly every step in the experimental process. A key finding was that the scope of needs goes beyond analysis of data generated from tests or experiment; integrated automated collection and analysis of process and instrument control data was also noted as very important, since this information is often needed to evaluate test data and results. In terms of experimental data, the scope of needs for excellent automated analysis tools is vast, but particular needs for computing tools are seen in image analysis and quantification and spectral peak detection and quantification.

- **Challenges with analysis tools used today:** Participants report using a large variety of commercial software, and home-built software to accomplish automated data collection and analysis in their work. Several commercial analysis packages emerged as meeting many of the needs. However, most participants reported the need to either supplement or modify current software extensively, or to work around problems or gaps in current software capabilities. In addition, participants noted that the most capable software packages required extensive training for their employees to use it. In each case, this adds substantial costs – sometimes on the order of 100% of the original software cost. The biggest problem with automated data analysis is that it is not robust and precise enough to handle the variety of data generated from high-throughput experiments. Participants report common unacceptable software analysis when the package was faced with unexpected data trends, and even when trends were more gradual. Indeed, one participant reported that if it was "important," they would still employ human operation in order to ensure proper data handing and analysis. In all cases, participants reported problems with integrating the analysis software with larger systems.
- **Software Needs and Opportunities:** Robust automating image analysis was most identified as a need for participant companies, and seems to represent a key opportunity for software vendors. Participants suggest that the types of image analysis routines (i.e., the distilled results they produce) are adequate, so there are not needs for new kinds of routines. However, the robustness of routines must be improved if they are to be applied in an automated manner to large streams of image data. Since the identification of trends is often most important, the quantification is not as important as robustness and repeatability. The second priority is automated spectral analysis. Here, routines that handle both systematic and unexpected drifts in peak location and magnitude are needed. The major needs to be met are similar to image analysis – robustness is the key. However, as opposed to image analysis, the analysis of spectra needs to be quantitative. Visualization software is the final priority. Here, some good software exists, but it tends to combine many kinds of visualization analysis in a single expensive package. Since most practitioners of high-throughput need to apply a single kind of visualization many times, it would be useful if routines could be purchases piece-by-piece.
- **Data Formatting Needs and Opportunities:** Most participant companies report that they would pay a premium for software and instrument output that was open source, or that had key data and systems integration capabilities included. Open source data output is extremely important, and a company may choose a software package based on whether the data format is proprietary, or poses other barriers to system integration. Some key points:
    - It is paramount that data refinement and reduction routines remain transparent and well-documented in software packages. Indeed, most companies note problems in software packages that perform data "massaging" operations that are hidden or not described fully, and note a related lack of confidence in these routines.
    - Related to the previous point, the ability access raw data from instrumentation is very important for integration, and for applying innovative, custom-built analysis tools. Many companies report that they would share analysis advances with instrument manufactures in return for access to raw data streams.
    - Participant companies urge software and instrument developers to export data into open source and non-proprietary formats. This would be useful if the developers

adopted emerging common formats, such as XML. In addition, it is noted that data outputs that include labeled, and fully described, instrument and software metadata (instrument operation parameters, software version, software operation parameters, etc.), considerably eased both system integration and data archiving – both of which are more important in high-throughput operation.

- **Datamining Needs and Opportunities:** For most companies, datamining needs are relatively basic and well defined. Simple searches for descriptors, value thresholds, value ranges, with some means to cross-correlate these over a few parameters usually suffice most needs. However, the lack of common data formats, pervasive closed data formats, and inadequate data descriptors seriously inhibit participant's ability to apply these simple datamining tools. Accordingly, participants urge developers towards instruments and software that produce open, documented data files as described above.

Session #3: Design of Experiment Tools and Lab Notebooks
Moderator: Christopher Stafford, NCMC, Polymers Division, NIST
Notes: Adam Nolte, NCMC, Polymers Division, NIST

Guiding Questions
Practitioners:

a) Do you use computer aided Design of Experiment (DOE) tools as part of your high-throughput experiment workflow?
b) Do current software tools enable what you need from DOE? If not, what capabilities would you like to see from a DOE software package?
c) Do you use an electronic Lab Notebook in your work? If so, what parts of workflow information to you store in it?
d) What capabilities do you need from a DOE package, and how would it improve your work?

Suppliers:

a) Does your instrument/software package include DOE or Lab Notebook functions?
b) What are the main barriers to including DOE functions in your instrument/software package? Unsure of user needs? Lack of market?

Discussion Summary

- Design of Experiment (DOE) Tools. Each company participating in the discussion used DOE tools in varying degrees. A number of software packages were used, with *Design-Expert*, *JMP*, and *SAS* being popular responses. At least one participant company uses an internally designed product. The general consensus was that DOE tools are useful and widely used, but could be improved in the following ways:
  - Clearer knowledge of which programs to use for which types of experiments. DOE encompasses many types of experimental design methodologies (e.g., screening experiments, optimization, failure analysis, etc.), each of which implies a different design strategy. There needs to be a clearer knowledge of which types of software are best for answering which DOE needs.
  - Greater flexibility in data transfer. Many companies expressed the need for easier transfer of experimental designs into the experimental workflow. This need encompasses both help in translating theoretical "lows" and "highs" in a design to physical values in experiments, as well as an ability to directly interface automated equipment with DOE tools so that experiment designs can be immediately implemented as physical experiments, without the need for operator intervention and manual entry of DOE parameters.
  - Greater flexibility in experiment design. Several companies expressed frustration with the difficulty of incorporating operator knowledge of experimental constraints into DOE planning. A scientist might know, for example, to avoid a certain concentration or temperature range, or that a more detailed phase diagram could be obtained by spacing experiments non-evenly in parameter space. Incorporating these types of restraints into current DOE software is apparently difficult or non-

intuitive, and probably reflects both the need for better training in DOE implementation and software flexibility.

- Electronic Lab Notebooks (ELN). ELNs are on every company's radar screen, but their current benefit is questioned by many. Key champions of this technology have been the healthcare/pharmaceutical industries, where ELNs allow for ease of documenting regulation compliance, and interfacing with laboratory information management systems (LIMS). It was generally agreed upon that ELNs should and will go hand-in-hand with the development of LIMS in the future, but presently most companies aren't convinced of the cost-benefit analysis to working with ELNs. The major concerns/objections to ELN technology involved the following:
  - Psychological attachment to paper. Many individuals voiced preference for the way a physical notebook looks, feels, and is used. Beyond simple psychological aspects, which many admitted could be overcome, there were still hard questions about the "real" purpose of ELNs. Many people expressed confusion about what their ultimate purpose was, as they are ultimately attempting to introduce change into a system (paper notebooks) that works well already. Companies need to be convinced that ELNs will introduce positive changes in areas such as IP protection, regulatory compliance, research collaboration, and productivity.
  - ELN function and performance. Some individuals expressed frustration that while ELNs worked well for tabulated data, they could not easily search through appended documents (PDFs, image files). Concerns were also raised about the ability of ELNs to be as universally useful as conventional notebooks. They seem now to be functionally marketed towards particular industrial needs, e.g. regulatory compliance or synthetic chemistry. Can they universally meet the diverse needs of scientists?
  - Legal concerns. There is a sense of legal confusion among companies about whether ELNs will be as equally admissible as their paper counterparts should legal disputes over IP arise.